Tracking The Evolution of Literary Style Via Dirichlet-Multinomial Change Point Regression

Tracking The Evolution of Literary Style Via Dirichlet-Multinomial Change Point Regression

Gordon J. Ross

University of Edinburgh, UK

E-mail: gordon@gordonjross.co.uk

Summary. It is typical in stylometry to assume that authors have a unique writing style which is common to all their published writings and constant over time. Based on this assumption, statistical techniques can be used to answer literary questions, such as authorship attribution, in a quantitative manner. However the claim that authors do indeed have a constant literary style has not received much investigation or validation. We propose a collection of statistical models based on Dirichlet Multinomial change point regression which are able to capture the evolution of writing style over time, including both gradual changes in style as the author matures, and abrupt changes which can be caused by extreme events in the author's life. To illustrate our framework, we study the literary output of the celebrated British author Sir Terry Pratchett, who was tragically diagnosed with Alzheimer's disease during the last years of life. Contrary to the usual assumptions made in stylometry, we find evidence of both gradual changes in style over his lifetime, and an abrupt change which corresponds to his Alzheimer's diagnosis. We also investigate the published writings of Agatha Christie who is also rumoured to have suffered from Alheizmer's towards the end of her life, and find evidence of gradual drift, but no corresponding abrupt change. The implications for stylometry and authorship attribution are discussed.

1. Introduction

Stylometry involves the quantification of linguistic style for the purpose of answering literary questions such as authorship attribution. The usual assumption is that authors have a distinctive writing style which is stable over time, and which can be identified in all their writings. As such, a statistical model that has been learned for the style of a particular author can be used to answer questions such as whether texts with disputed authorship were written by the author in question. A classic example of authorship attribution comes from Mosteller and Wallace (1963) who used grammatical features such as function word frequencies and word length distributions to determine whether twelve disputed essays in *The Federalist Papers* were written by James Madison, or Alexander Hamilton. Other typical examples include Thisted and Efron (1987) where an alleged newly discovered Shakespeare poem is tested for authenticity, and Abakuks (2012) which uses statistical techniques to investigate the relationship between the various authors of the Synoptic Gospels. Reviews of the related stylometry literature can be found in Juola (2006), Holmes (1985) and Peng and Hengartner (2002).

In addition to the traditional literary problem of authorship attribution, the quantification of writing style is becoming increasingly important in fields such as cybersecurity where identifying the authors of anonymous internet forum posts can be an integral part of forensic investigation (Pearl and Steyvers, 2012; Narayanan et al., 2012), and plagiarism detection (Uzuner et al., 2005; Lukashenko et al., 2007). However, a potential limitation of most statistical approaches to stylometry is the (often unstated) assumption that the style of an author remains constant over time, and does not undergo change. In authorship attribution tasks where the goal is to determine which author from a candidate set wrote a particular disputed work, the usual procedure is to assemble a corpus of texts known to be written by each author, and compare the similarity of the disputed work to quantitative features extracted from those corpuses (Koppel et al., 2007; Madigan et al., 2005). However this procedure implicitly assumes that different texts written by the same author have an identical

literary style. This is somewhat questionable, since it may be the case that an author's style gradually evolves over time as their writing matures, or changes abruptly in response to events in the author's life.

Despite its importance, the extent to which the literary style of an author can change over time has received little attention. One exception is Can and Patton (2004) who examine the body of work of two Turkish authors. For the first author, two corpuses of works are assembled, the first of which contains the newspapers articles that he wrote during the period 1960-1969, and the second of which contains the newspaper articles that he wrote in 2000. A variety of tests are used to conclude that the literary style in both corpuses is different, suggesting that style can change over time, and similar results are obtained for the second author. However, knowing that linguistic style can change between corpuses of work written decades apart provides only limited information about the evolution of an author's style over shorter periods of time. More recently, Hirst and Feng (2012) and Le et al. (2011) have used stylometric techniques for the purpose of detecting Alzheimer's disease and dementia in authors. Their hypothesis is that the onset of these diseases causes cognitive impairment which can manifest as changes in writing style. As in Can and Patton (2004) the corpus of a small number of authors was split into different periods, and tests were applied to check whether the writing style in both was identical, with a difference taken to be evidence of possible cognitive impairment. However this suffers from the limitation that it cannot distinguish between a single abrupt change due to cognitive decay, and the cumulative effect of gradual change over time due to natural evolution as the writer's style matures.

The current article develops new statistical methodology for modelling potential changes in the style of an author (or group of authors) over time. Intuitively, we can view the process of stylistic change as having two potential forms. The first is the gradual evolution of style that takes place as an author matures, and which would be expected to be slow moving. The second corresponds to abrupt changes which may produce a very fast alteration in style over a short period of time, perhaps in response to extreme events in the author's life. Our methodology consists of representing each text in an author's corpus as a product of compound Dirichlet-Multinomial distributions, and using a change point regression framework which allows gradual change to appear as a drift term, with change points corresponding to abrupt stylistic changes. To our knowledge, this is the first time that the Dirichlet-Multinomial distribution has been used in stylometry, although Giron et al. (2005) speculated about its applicability, and it has seen some usage in the related field of document classification (Madsen et al., 2005; Doyle and Elkan, 2009) without regard for potential changes over time. We will show later that it corrects many of the problems that are associated with the Multinomial distribution which is common in the stylometry literature (Giron et al., 2005; Riba and Ginebra, 2006).

We illustrate our methodology through a study of the literary output of two celebrated authors. The bulk of the paper focuses on the British author Sir Terry Pratchett, best known for writing the popular "Discworld" series of fantasy novels. Pratchett is an interesting case study for three reasons. First, he has a large literary output consisting of over 40 novels written over a 30 year period, which is a long enough time horizon to expect stylistic changes to occur. Second, the majority of his fictional works were set in the same genre, which avoids the well-known problem that books written by the same author but in different genres can have markedly different styles (Smith, 1983). Finally, during the later years of his life, Pratchett was tragically diagnosed with Alzheimers disease (BBC News, 2015) which impaired his ability to write, causing him to partially switch to dictating his books to an assistant. We will show that this produced an identifiable abrupt change in the literary style of his

later novels, which our methodology is able to detect. Unlike Hirst and Feng (2012) and Le et al. (2011), we are able to distinguish this abrupt change due to Alzheimer's from gradual stylistic evolution, and we show that the change occurs immediately after the Alzheimer's diagnosis. To illustrate the robustness of our methods, we also briefly analyse the literary output of another prominent author – Dame Agatha Christie – who is also rumoured (although not known) to have contracted Alzheimer's towards the end of her life. Unlike the case of Pratchett, we do not find evidence of an abrupt change in the later writings of Christie.

The remainder of this article proceeds as follows. In Section 2 we review the features typically used to characterise literary style, and describe the Pratchett corpus in more detail. Exploratory data analysis of this corpus is used to show that the assumption of constant style over time is doubtful. Next in Section 3 we introduce a number of models for literary style which can incorporate both gradual and abrupt changes. These models are applied to the Pratchett and Christie corpuses in Section 4, where it is shown that model selection strongly favours an approach that can features both gradual and abrupt change. Finally the implications for both authorship attribution and stylometry are discussed in Section 5.

2. Background and Data

2.1. Quantitative Description of Texts

Let $\mathbf{B} = (B_1, \ldots, B_n)$ be a time-ordered corpus of texts (such as books or journal articles) written by a particular author, listed in time-order of publication so that B_1 is the text that was written first, B_2 is the text that was written second, and so on. Each text B_i consists of N_i words, and we write $w_{i,j}$ to denote the j^{th} word in text i. The commonly used bag-of-words model (Wallace, 2006) assumes that the author has access to a vocabulary containing an unknown number v of words, and that each word in the text is independently drawn from this vocabulary at random. Each word $w_{i,j}$ can hence be viewed as a draw from a text-specific Multinomial distribution with probability vector $\boldsymbol{\pi}_i = (\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,v})$ defined over the author's vocabulary. A word-frequency vector $\mathbf{c}_i = (c_{i,1}, \ldots, c_{i,v})$ can then be associated with each text, where $c_{i,k}$ denotes the number of times the k^{th} word in the author's vocabulary appeared in text i and $\mathbf{c}_i \sim \text{Multinomial}(N_i, \boldsymbol{\pi}_i)$. In almost all existing studies in stylometry and authorship attribution, the author's style is assumed to be constant over time so that $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 = \ldots = \boldsymbol{\pi}_n$.

A drawback of using naive word frequency vectors to describe the writing style of an author is that the word frequencies can depend strongly on the semantic content of the text. For example, in a novel set in the city of Paris, the word 'Paris' will occur much more frequently than it would in a typical novel, however this is not particularly indicative of the author's literary style. As such, it is typical in stylometric to instead define writing style using features which are independent of meaning or context, since these are less subject to the conscious control of the author and hence more likely to be stable across texts (Holmes, 1985; Narayanan et al., 2012). A wide variety of features have been considered in the literature, with Koppel et al. (2009) giving a recent review. One of the most commonly used features is the frequency count of function words in a text,

The term 'function words' is used to describe words which have little independent meaning and are mainly used as part of the grammar of the English language. Examples include words such as 'a', 'are', 'but', 'to', and 'if'. These words are expected to be present in all texts of sufficient length, and their frequencies should not fluctuate much based on semantic context. The modern usage of function words in stylometry is discussed extensively in Argamon and Levitan (2005) and Zhao and Zobel (2005). A recent trend in stylometry has been to expand the definition of function word slightly, to include the most commonly used words in a typical

the, and, to, a, of, I, in, he, was, it, you, that, his, said, her, she, had, with, as, not, but, at, for, on, is, be, have, him, they, all, were, what, me, there, one, my, this, if, no, from, so, would, up, out, by, been, them, or, do, could, when, we, an, are, who, like, know, about, your, did, mr, then, will, very, its, into, their, now, more, man, well, down, which, some, see, back, time, think, just, than, dont, little, can, only, here, any, way, how, over, thought, other, good, say, never, too, looked, much, before, come, two, go, again, old, even, has, made, might, where, head, right, eyes, got, after, mrs, still, yes, hand, off, something, face, should, away, through, must, people, sir, get, though, miss, look, long, us, came, going, went, am, himself, make, why, men, own, big, around, im, those, take, lord, seemed, first, tell, being, always, another, quite, woman, upon, want, things, nothing, last, door, these, didnt, such, oh, knew, once, took, great, really, put, thing, day, young, told, voice, our, let, most, enough, thats, because, every, room, turned, may, left, without, saw, many, course, anything, looking, ever, asked, heard, yet, night, find, done

Fig. 1. List of the 200 most common words that we use to characterise literary style.

,

text. We follow this approach and extract the 200 most commonly used words in the Pratchett corpus, which we will use to define literary style. These 200 common words are shown in Table 1 in decreasing order of frequency count. It can be seen that the most commonly used words correspond to traditional function words ('the', 'and', 'to') while the list also includes some non-grammatical words which are nonetheless very widely used in English in a relatively context-independent manner ('without', 'anything', 'because'). For the rest of this paper, we will use the term 'function words' to refer to these 200 most common words.

We hence associate a 201 element vector $\mathbf{c}_i = (c_{i,1}, \ldots, c_{i,201})$ with each text B_i where for $1 \leq j \leq 200$, the variable $c_{i,j}$ counts the number of times that the j^{th} function word appears in the text, and $c_{i,201} = N_i - \sum_{j=1}^{200} c_{i,j}$ counts the number of non-function words in the text. Under the bag-of-words model, the $\mathbf{c}_i$ vector has a Multinomial$(N_i, \boldsymbol{\pi}_i)$ distribution where $\pi_{i,j}$ is the probability of a randomly selected word in text i being the j^{th} function word. To keep our later notation more general, we will write $W = 201$ to refer to the length of the function words vectors rather than using the specific number (201).

2.2. The Discworld Corpus

The main literary corpus we analyse contains the Discworld novels written by the celebrated British author Terry Pratchett. The Discworld series consists of 41 fantasy novels written between the years 1983 and 2015, and a full list of books along with publication year is given in Table 9 in the Appendix. As discussed in the introduction, this corpus is ideal for studying the evolution of literary style over time, since there is strong a priori reason to believe that it might contain both gradual change due to the length of the writing period, and abrupt change due to the onset of Alzheimer's disease.

For each novel in the corpus, we produced a digitalization using Optical Character Recognition, and extracted the vectors corresponding to the counts of function words. As discussed above, each novel B_i is then represented as a length 201 vector describing the frequency of the function words.

Before beginning our formal modelling, we first use a type of Multidimensional Scaling (Borg and Groenen, 2005) to visualise the corpus and provide motivation for a model which allows for both gradual and abrupt changes in writing style. For each book in the corpus, we first normalise its function word frequency to sum to 1, and then standardise each of

Fig. 2. Linguistic features for each of the 41 numbered Discworld novel projected onto the first and second principal components of the (normalized) function word frequencies. Young adult books are colored in red

the 201 features so that each has mean 0 and standard deviation 1. The Euclidean distance between all pairs of books is then computed, and Multidimensional Scaling is then used to find the best two dimensional projection of the resulting distance matrix (note that although the word frequencies are discrete rather than continuous, the wide range of each feature means that treating them as continuous is not unreasonable). Figure 2 shows the resulting 2-d representation, where each novel is represented by its number in the corpus (i.e. the first novel B_1 in the Discworld series is shown as the number '1', and so on). The distance between each pair of novels on the plot roughly corresponds to how similar they are in linguistic style, with more similar books being closer together. Several interesting features can be observed in this plot:

- The linguistic style seems to be gradually changing over time, as demonstrated by the fact that books which were written around the same time tend to be closer together than books which were written further apart.

- There is preliminary evidence of a change point that occurs somewhere around book number 36 (titled 'Making Money'), which was published in September 2007. Terry Pratchett

publicly announced his Alzheimers diagnosis in December 2007, with the next book 38 being published in 2008. As such, this suggests that the hypothesis that Alzheimer's caused a change in writing style may be accurate.

- Six of the books in the Discworld series were written specifically for 'Young Adults', and hence form a subcorpus within the main corpus. These books are colored red on the plot, and seem to have a slightly different linguistic style to the rest of the corpus, which were written for a general audience. The first four young adult books seem to form a cluster towards the top of the plot. This is consistent with a well known finding in stylometrics where books written by the same author but in different genres can have slightly different styles (Smith, 1983).

Next, Figure 3 shows how some of the individual features change over time in the Pratchett corpus. Each plot shows an extract from the function word proportion vectors $\mathbf{c}_i/N_i$ which shows the proportion of times the words 'the', 'to', 'that', 'on', 'if' and 'might' appeared in each of the 41 books. A diversity of patterns can seen for the different function words, with some appearing to undergo little change over time (e.g. 'on') while others undergo gradual drift which corresponds to the evolution of literary style and generally seems to take the form of a linear trend. Additionally, some words such as 'that', 'to' and 'might' seem to show evidence of a change point towards the end of the corpus, which again points towards stylistic change. Similar patterns can be observed in many of the other (unshown) function word features.

Based on the above analysis, we will analyse the corpus of Discworld novels with the Young Adult books removed. We also note in passing that Pratchett also wrote several other novels that are not part of the Discworld series. We exclude these from our analysis to avoid the possibility that these have fundamentally different styles to the Discworld novels.

3. Methodology

We now introduce a collection of statistical models for describing the literary style of authors, which takes into account potential changes over time. In Section 4 we will fit all these models to the Pratchett corpus and show how model selection can be used to choose the most appropriate one, based on how the style of the author evolves.

We begin with the most simple model which assumes that each book in the corpus can be viewed as an independent draw from a Multinomial distribution with fixed parameters. This is the predominant methodology used in the parts of the stylometric literature which use explicit statistical models (Gill and Swartz, 2011; Riba and Ginebra, 2006; Giron et al., 2005). However, in cases where the stylistic features have high levels of variance across different texts in the corpus, the Multinomial model will be underdispersed. As such, we also introduce the Dirichlet-Multinomial compound distribution which allows for over-dispersion. Next, we consider models that take time-evolution into account, starting with gradual drift, followed by abrupt change, and then a model which incorporates both.

3.1. Multinomial vs Dirichlet-Multinomial Modeling

It is common in the stylometrics literature to use the Multinomial distribution to model the function word counts (Gill and Swartz, 2011; Riba and Ginebra, 2006; Giron et al., 2005). In this case, the function word vector $\mathbf{c}_i$ associated with each book B_i in the corpus is represented as a draw from a Multinomial distribution with parameter $\boldsymbol{\pi}_i = (\pi_{i,1}, \ldots, \pi_{i,W})$. Therefore we have $B_i \sim \text{Multinomial}(\mathbf{c}_i; N_i, \boldsymbol{\pi}_i)$ with pdf:

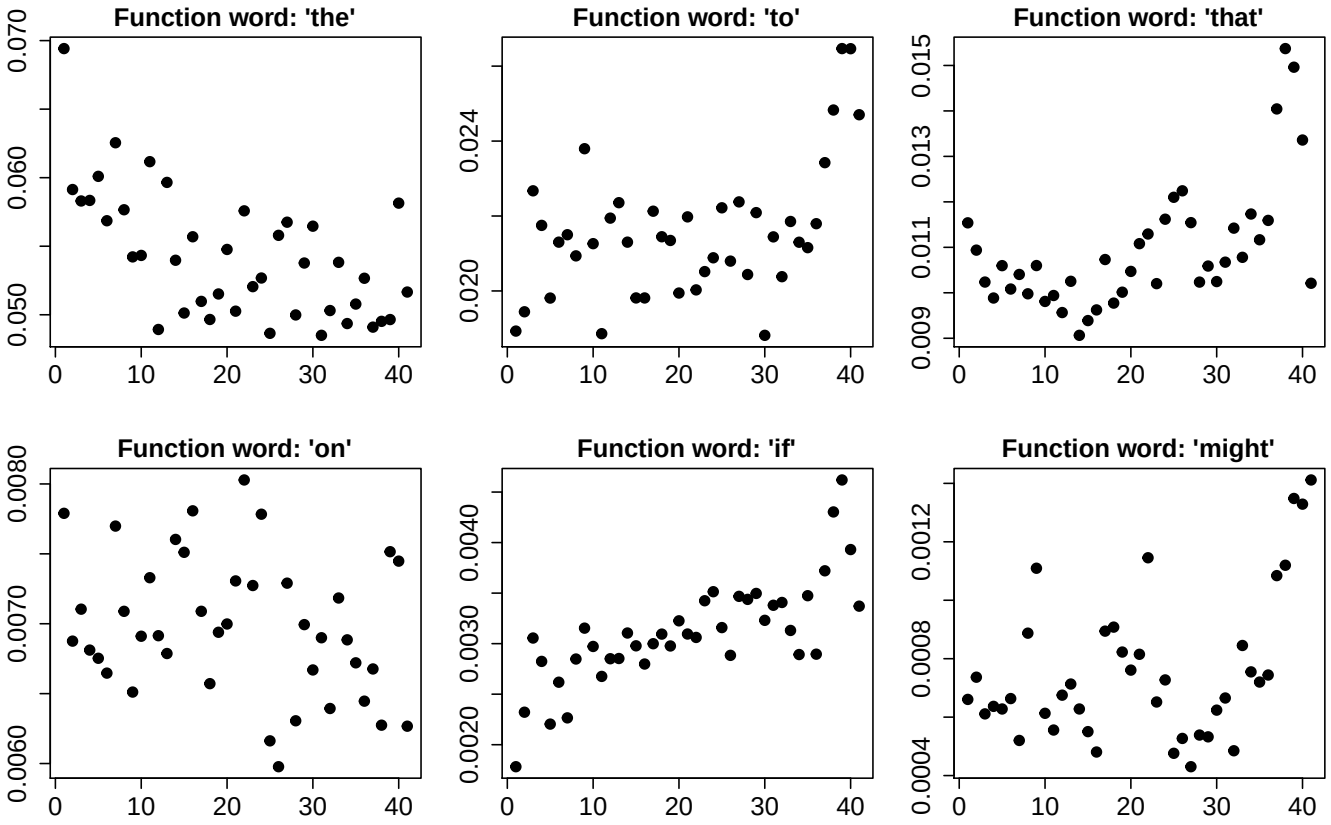

Fig. 3. Evolution of some selected function word frequencies over time in the Pratchett corpus. Each plot shows the proportion of words in each book which were the given function word.

$$p(B_i|\boldsymbol{\pi}_i) = \frac{N_i!}{c_{i,1}, \ldots, c_{i,W}} \prod_{j=1}^{W} (\pi_{i,j})^{c_{i,j}}$$

where N_i is the number of words in book i, $c_{i,j}$ is the number of times the j^{th} function word appears in book i, W is the number of function word features, and we have the constraints $\sum_{j=1}^{W} \pi_{i,j} = 1$. If we assume that there is no change in literary style over time, so that the $\boldsymbol{\pi}_i = \boldsymbol{\pi}$ are constant and independent of i. then the maximum likelihood estimates are:

$$\hat{\boldsymbol{\pi}}_{,j} = \frac{\sum_{i=1}^{n} c_{i,j}}{\sum_{i=1}^{n} N_i}$$

where $\hat{\boldsymbol{\pi}}_{,j}$ denotes the j^{th} component of the vector $\hat{\boldsymbol{\pi}}$ (i.e. the proportion associated with the j^{th} function word) and n is the number of books in the corpus. Note that we use the notation $\hat{\boldsymbol{\pi}}_{,j}$ to distinguish between the elements of the vector $\hat{\boldsymbol{\pi}}$ which is common to all books (under the assumption of no changes in style) and the book-specific $\hat{\boldsymbol{\pi}}_i$ vectors.

Although the Multinomial distribution is widely used in stylometry, there has been little investigation into whether it is actually an appropriate statistical model for literary corpuses. In our empirical work, we have found that most corpuses tend to have over-dispersion that cannot be well described by the Multinomial distribution. To assess the fit of the Multinomial distribution for the Pratchett corpus, we can use a test statistic based on the sum of squared standardised residuals (Pierce and Schafer, 1986):

$$S = \sum_{i=1}^{n} \left[\sum_{j=1}^{W} \left(\frac{c_{i,j} - \hat{\mu}_{i,j}}{\hat{\sigma}_{i,j}} \right)^2 \right] \tag{1}$$

where $\hat{\mu}_{i,j} = N_i \hat{\boldsymbol{\pi}}_{,j}$ and $\hat{\sigma}_{i,j} = \sqrt{N_i \hat{\boldsymbol{\pi}}_{,j}(1 - \hat{\boldsymbol{\pi}}_{,j})}$ are the estimated means and standard deviations of the function word counts in text i under the Multinomial distribution. The p-value for the test on the Pratchett corpus computed using a parametric bootstrap (Efron and Tibshirani, 1983) was less than 1×10^{-5} indicating a severe lack of fit. Similar results were observed on both the Christie corpus.

To deal with this potential over-dispersion, we also consider modelling the function word vectors using the compound Dirichlet-Multinomial (DirMult) distribution instead. In this case, each count vector $\mathbf{c}_i$ is a draw from a Dirihclet-Multinomial distribution with parameter $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \ldots, \alpha_{i,W})$. Therefore we have $B_i \sim \text{DirMult}(\mathbf{c}_i; N_i, \boldsymbol{\alpha}_i)$ with pdf:

$$p(B_i|\boldsymbol{\alpha}_i) = \frac{\Gamma(A_i)N!}{\Gamma(N_i + A_i)} \prod_{j=1}^{W} \frac{\Gamma(c_{i,j} + \alpha_{i,j})}{\Gamma(\alpha_{i,j})c_{i,j}!}, \quad A_i = \sum_{j=1}^{W} \alpha_{i,j}$$

The Dirichlet-Multinomial distribution is an example of a compound distribution, which arises from assuming that the $\boldsymbol{\pi}$ parameter in the Multinomial distribution is itself stochastic, and follows a Dirichlet distribution. Specifically, if we have that $\mathbf{c} \sim \text{Multinomial}(\boldsymbol{\pi})$ and $\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, then the marginal distribution of $\mathbf{c}$ is $\text{DirMult}(\boldsymbol{\alpha})$. Unlike the $\boldsymbol{\pi}$ parameter in the Multinomial distribution, the $\boldsymbol{\alpha}$ vector is not constrained to sum to one, and the additional parameter provides an extra degree of freedom which allows for over-dispersion. The Diricihlet-Multinomial distribution is commonly used in the document classification literature (Madsen et al., 2005; Doyle and Elkan, 2009) where it forms the basis of the Latent Dirichlet Allocation algorithm but its application to stylometry has not previously been investigated.

Again if we assume that there is no change in literary style over time, so that the $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}$ are constant and independent of i, then the maximum likelihood estimate can be found by maximising the corpus likelihood function:

$$L(\boldsymbol{\alpha}|\mathbf{B}) = \prod_{i=1}^{n} \left(\frac{\Gamma(A_i)}{\Gamma(N_i + A_i)} \prod_{j=1}^{W} \frac{\Gamma(c_{i,j} + \boldsymbol{\alpha}_{,j})}{\Gamma(\boldsymbol{\alpha}_{,j})} \right), \quad A_i = \sum_{j=1}^{W} \boldsymbol{\alpha}_{,j}$$

where $\boldsymbol{\alpha}_{,j}$ denotes the j^{th} component of the vector $\hat{\boldsymbol{\alpha}}$. Unlike in the Multinomial case, this likelihood function cannot be maximised analytically and so a closed form expression for $\hat{\boldsymbol{\alpha}}$ is not available. However, it is easy to maximise the likelihood numerically using either direct Newton-Raphson gradient ascent, or one of the alternative techniques described in Minka (2000).

Performing the same goodness-of-fit test using the score statistic from Equation 1 gave a p-value of 0.54 on the Pratchett corpus, indicating the DirMult distribution seens to give a reasonable fit to the corpus. Similar results were observed on the Christie corpus. To further compare the Multinomial and DirMult distributions, we can perform direct model selection using penalised likelihood with the standard Bayesian Information Criterion (BIC) penalty (Schwarz, 1978), which selects the model which maximises the BIC function:

$$BIC(\mathbf{B}) = \log p(\mathbf{B}|\hat{\boldsymbol{\theta}}) - \frac{k}{2} \log(n),$$

where $\boldsymbol{\theta}$ denotes the parameter vector for the model with corresponding maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, k is the number of parameters in the model, and n is the number of texts

in the corpus. For the Multinomial distribution we have $\boldsymbol{\theta} = (\boldsymbol{\pi})$ and $k = W - 1$, while for the DirMult distribution we have $\boldsymbol{\theta} = (\boldsymbol{\alpha})$ and $k = W$. The additional parameter in the DirMult come from the lack of sum-to-one constraint in the parameter vector. We shall see in Section 4 that the Dirichlet-Multinomial distribution gives a substantially superior fit to all three corpuses.

3.2. Gradual Drift

We now consider a model specification where the author's literary style can change gradually over time as their writing matures. In this case. the $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ parameters governing the function word distributions become functions of time. Motivated by the previous discussion in Section 2 and Figure 3, we model this type of parameter drift using a set logistic regression models with a linear drift specification (Chen and Li, 2013):, which in the case of the Dirichlet-Multinomial distribution can be written as

$$\alpha_{i,j} = \exp(\beta_j^0 + i\beta_j^1) \tag{2}$$

where $\alpha_{i,j}$ again denotes the j^{th} element of $\boldsymbol{\alpha}_i$ for the i^{th} book, and the β variables are word specific regression coefficients. In other words, each of the W function word features are assumed to follow a linear regression, resulting in 2W overall parameters.

The drift for the Multinomial distribution is also specified using a set of logistic regression models, with the added constraint that the $\boldsymbol{\pi}_i$ vectors must sum to 1 for all i. This can be enforced by using the specification (Menard, 2002)

$$\pi_{i,j} = \begin{cases} \frac{\exp(\beta_j^0 + i\beta_j^1)}{1+\sum_{j=1}^{W-1}\exp(\beta_j^0 + i\beta_j^1)} & \text{if } j < W \\ \frac{1}{1+\sum_{j=1}^{W-1}\exp(\beta_j^0 + i\beta_j^1)} & \text{if } j = W \end{cases} \tag{3}$$

Fitting either of these models to the corpus requires estimation of the $\boldsymbol{\beta} = (\beta_1^0, \beta_1^1, \ldots, \beta_W^0, \beta_W^1)$ coefficients. In neither the Multinomial nor the Dirichlet-Multinomial case are the maximum likelihood estimates available in closed form, however numerical techniques such as Newton-Raphson can easily be used to maximise the corresponding log-likelihood function of the corpus. The log-likelihood function for the Dirichlet-Multinomial regression model is (Chen and Li, 2013):

$$\log L(\boldsymbol{\beta}|\mathbf{B})_{DirMult} \propto \sum_{i=1}^{n} \left[\tilde{\Gamma}\left(\sum_{j=1}^{W} \exp(\beta_j^0 + i\beta_j^1)\right) - \tilde{\Gamma}\left(\sum_{j=1}^{W} c_{i,j} + \exp(\beta_j^0 + i\beta_j^1)\right) + \right.$$
$$\left. + \sum_{j=1}^{W} \tilde{\Gamma}(c_{i,j} + \exp(\beta_j^0 + i\beta_j^1)) - \tilde{\Gamma}(\exp(\beta_j^0 + i\beta_j^1)) \right] \tag{4}$$

where $\tilde{\Gamma}$ denotes the logarithm of the Gamma function. For the Multinomial distribution, the loglikelihood function is (Agresti, 2013):

$$\log L(\boldsymbol{\beta}|\mathbf{B})_{Mult} \propto \sum_{i=1}^{n} \left[\sum_{j=1}^{W-1} c_{i,j}(\beta_j^0 + i\beta_j^1) - \log\left(1 + \sum_{j=1}^{W-1} \exp(\beta_j^0 + i\beta_j^1)\right) \right] \tag{5}$$

3.3. Abrupt Changes In Literary Style

We next consider the case where literary style can undergo abrupt change, such as in the potential case of Alzheimers. The simplest situation is when abrupt changes are the only possible source of stylistic evolution, i.e. there is no gradual drift. Suppose that there are K change points $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ where each change may correspond to a shift in the distributional parameters. In the case of the Dirichlet-Multinomial distribution, the probability model is then:

$$p(B_i = \mathbf{c}_i|\boldsymbol{\tau}, \boldsymbol{\alpha}) \sim \begin{cases} \text{DirMult}(\mathbf{c}_i; N_i, \boldsymbol{\alpha}_0) & \text{if } i < \tau_1 \\ \text{DirMult}(\mathbf{c}_i; N_i, \boldsymbol{\alpha}_1) & \text{if } \tau_1 \leq i < \tau_2 \\ \text{DirMult}(\mathbf{c}_i; N_i, \boldsymbol{\alpha}_2) & \text{if } \tau_2 \leq i < \tau_3 \\ \ldots & \\ \text{DirMult}(\mathbf{c}_i; N_i, \boldsymbol{\alpha}_K) & \text{if } \tau_K \leq i < n \end{cases}$$

where the segment-specific $\boldsymbol{\alpha}_k$ parameters denote the $K+1$ values of $\boldsymbol{\alpha}$. For each change point τ_j, we have that $\boldsymbol{\alpha}_{j-1} \neq \boldsymbol{\alpha}_j$. Fitting this model to data requires estimating both the number of change points K, their locations $\tau_1, \ldots, \tau_K$, and the Dirichlet-Multinomial parameters $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0, \ldots, \boldsymbol{\alpha}_K)$. A change point model for the Multinomial distribution can be specified in a similar way.

Multiple change point models of this form have been widely studied in the statistics literature (Killick et al., 2013; Green, 1995; Ross et al., 2011), although we are not aware of any work which specifically relates to the Dirichlet-Multinomial distribution. To estimate the unknown parameters $\boldsymbol{\theta} = (K, \boldsymbol{\tau}, \boldsymbol{\alpha})$ we use a penalized maximum likelihood approach with the Bayesian Information Criterion (BIC) penalty previously discussed in Section 3.1, which is a relatively standard approach to parametric change detection (Killick et al., 2013) and involves maximising the penalised change point likelihood function:

$$BIC(\mathbf{B}) = \log p(B_{1:\tau_1}|\hat{\boldsymbol{\alpha}}_0) + \left(\sum_{i=2}^{K} \log p(B_{(\tau_{i-1}+1):\tau_i}|\hat{\boldsymbol{\alpha}}_i) \right) + \log p(B_{(\tau_K+1):n}|\hat{\boldsymbol{\alpha}}_K) \; - 0.5k \log(n) \tag{6}$$

where n is the number of books in the corpus, and k is the number of free parameters in the model. We use the notation $B_{r:s}$ to denote the set of books $(B_r, B_{r+1}, \ldots, B_s)$. For the Pratchett corpus with K change points, we have $n = 35$ (since the Young Adult books are omitted) and $k = W(K+1)+K$ for the Dirichlet-Multinomial model, and $k = (W-1)(K+1)+K$ for the Multinomial model. The (penalised) maximum likelihood estimates $\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\alpha}}$ then come from maximising Equation 6 over the parameter space. Naive maximization of this function is computationally demanding since there are 2^n possible change point configurations. To circumvent this problem, we can use a variant of the Pruned Exact Linear Time (PELT) algorithm introduced in Killick et al. (2013) which uses dynamic programming to perform the optimisation with a time-complexity between $O(n^2)$ and $O(n)$. We will now provide a brief overview of this algorithm, but for more details see Killick et al. (2013). Let $LL(B_{1:n})$ denote the log-likelihood of books $B_1, \ldots, B_n$ and let $\beta = 0.5k \log(n)$ be the BIC penalty for each additional change point. Choosing the number of change points K is equivalent to maximising the expression:

$$\sum_{i=1}^{K} \left(LL(B_{(\tau_{i-1}+1):\tau_i}) - \beta \right) \tag{7}$$

Suppose that we knew the configuration of change points in a subset $B_{1:s}$ of the data, and let τ^* denote the most recent one. Then the maximised penalised likelihood of the whole data is equal to the sum of the log-likelihood up until τ^* which was assumed to be known (since the configuration is known) and the log-likelihood of the data from $\tau^* + 1$ onwards, which contains an unknown number of change points. This leads to the following recursion: let $F(s)$ denote the maximised value of Equation 7 on the data $B_{1:s}$ and let $\mathcal{T}_s$ be all possible configurations of change points on this subset . Then:

$$\begin{aligned} F(s) &= \max_{\tau \in \mathcal{T}_s} \left[\sum_{i=1}^{K} \left(LL(B_{(\tau_{i-1}+1):\tau_i}) - \beta \right) \right] \\ &= \max_{t} \left(F(t) + LL(B_{(t+1):n}) + \beta \right), \quad t < s \end{aligned}$$

This recursion allows the $F(1), F(2), \ldots, F(n)$ values to be computed sequentially using dynamic programming, with $F(n)$ corresponding to the penalised maximised likelihood over the whole data-set. This reduces the task of finding change-points to an $O(n^2)$ operation. A further improvement to $O(n)$ can be achieved by pruning possible change-point configurations. For more details, see Killick et al. (2013).

Note that while the BIC penalty is widely used for selecting the number of change points in multiple change point models, several other methods for this purpose have also been proposed. This includes adjusments to the BIC which take into account the change point locations as well as their number K (Chen et al., 2006; Pan and Chen, 2006), however we found that these adjustments do not have much effect since the amount of penalisation in our formulation is dominated by the high number of model parameters (200). Since the BIC is derived as an asymptotic approximation to a Bayesian marginal likelihood, several higher order corrections have also been proposed (Hannart and Naveau, 2012; Zhang and Siegmund, 2007) along with fully Bayesian approaches (Green, 1995; Fearnhead, 2006). There are also alternative formulations of the multiple change point problem which do not rely on penalised likelihoods, such as Wild Binary Segmentation (Fryzlewicz, 2014). However as we will show in Section 4.2 through a simulation study, the standard BIC penalty is appropriate for our task here since it is able to adequately detect change points while also not producing a substantial number of false positive detections.

3.4. Gradual Drift and Abrupt Changes

Finally we introduce models combining both gradual drift and abrupt change, which can take into account both slow-changing stylistic evolution over time, as well as more rapid changes. We use a change point logistic regression model under which the elements of the function word parameter vector $\boldsymbol{\beta}$ under the Dirichlet-Multinomial specification evolve over time as:

$$\alpha_{i,j} = \begin{cases} \exp(\beta^0_{j,0} + i\beta^1_{j,0}), & \text{if } i < \tau_1 \\ \exp(\beta^0_{j,1} + i\beta^1_{j,1}), & \text{if } \tau_1 \leq i < \tau_2 \\ \exp(\beta^0_{j,2} + i\beta^1_{j,2}), & \text{if } \tau_2 \leq i < \tau_3 \\ \qquad\qquad \ldots & \\ \exp(\beta^0_{j,K} + i\beta^1_{j,K}), & \text{if } \tau_K \leq i < n \end{cases}$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ is a vector of K change points as in Section 3.3. This is an extension of the previous Equation 2 for gradual drift, which allows the coefficients to undergo abrupt change. As before, the number and location of the change points can be estimated along with the regression coefficients by minimising the likelihood penalised by the BIC, where there are

Model	Multinomial	Dirichlet-Multinomial
Constant	-68024	-35875
Gradual Drift	-60224	-35554
Abrupt Changes	-51635	-35682
Drift + Changes	-45319	-35541

Fig. 4. Maximum likelihood of model fits on the Pratchett corpus, penalised by BIC penalty. Larger (less negative) values indicate a more preferable model.

Model	Multinomial	Dirichlet-Multinomial
Abrupt Changes	7	1
Drift + Changes	7	1

Fig. 5. Number of change points found by each model.

now $k = 2W(K-1) + K$ parameters. A similar model can be defined using the Multinomial distribution by allowing the coefficients from the regression in Equation 3 to undergo change, in which case $k = 2(W-1)(K-1) + K$.

4. Results

4.1. Pratchett Corpus

To summarise the above, we consider the following four models to describe the literary style of a given corpus:

(a) A baseline model which assumes constant literary style where the parameters governing the function words distribution do not change over time. This is the approach used in almost all of the existing stylometry literature.
(b) Gradual drift where the linguistic style undergoes linear change over time, corresponding to small changes in style as the author's writing matures.
(c) A change point model where the linguistic style can undergo abrupt change, which may correspond to events such as the author switching genre, or undergoing a major life change.
(d) A combined model which allows for both gradual drift, and abrupt changes.

For each of these potential types of time-variation, we consider both the Multinomial specification that is common in stylometry, along with the Dirichlet-Multinomial specification discussed in Section 3.1. As such, we are comparing 8 different models in total.

In order to perform the comparison , we use penalised likelihood approach based on the BIC penalty. Table 4 shows the resulting BICs from fitting all eight models to the Pratchett corpus, and 5 shows the resulting number of change points in each model. Note that we removed the young adult novels when performing the comparison for reasons previously discussed in Section 2. The following results can be observed:

- The Dirichlet-Multinomial distribution gives a substantially better fit to the corpus than the Multinomial,, regardless of any considerations about time-variation. This is due to the variance of the frequency of function words counts over the corpus being substantially higher than predicted by the Multinomial. The compound distribution allows for this over dispersion due to the additional hierarchical layer.

- Although the models which allow for gradual drift and abrupt change separately give substantially better fits than the baseline model, the best fit is given by the model which

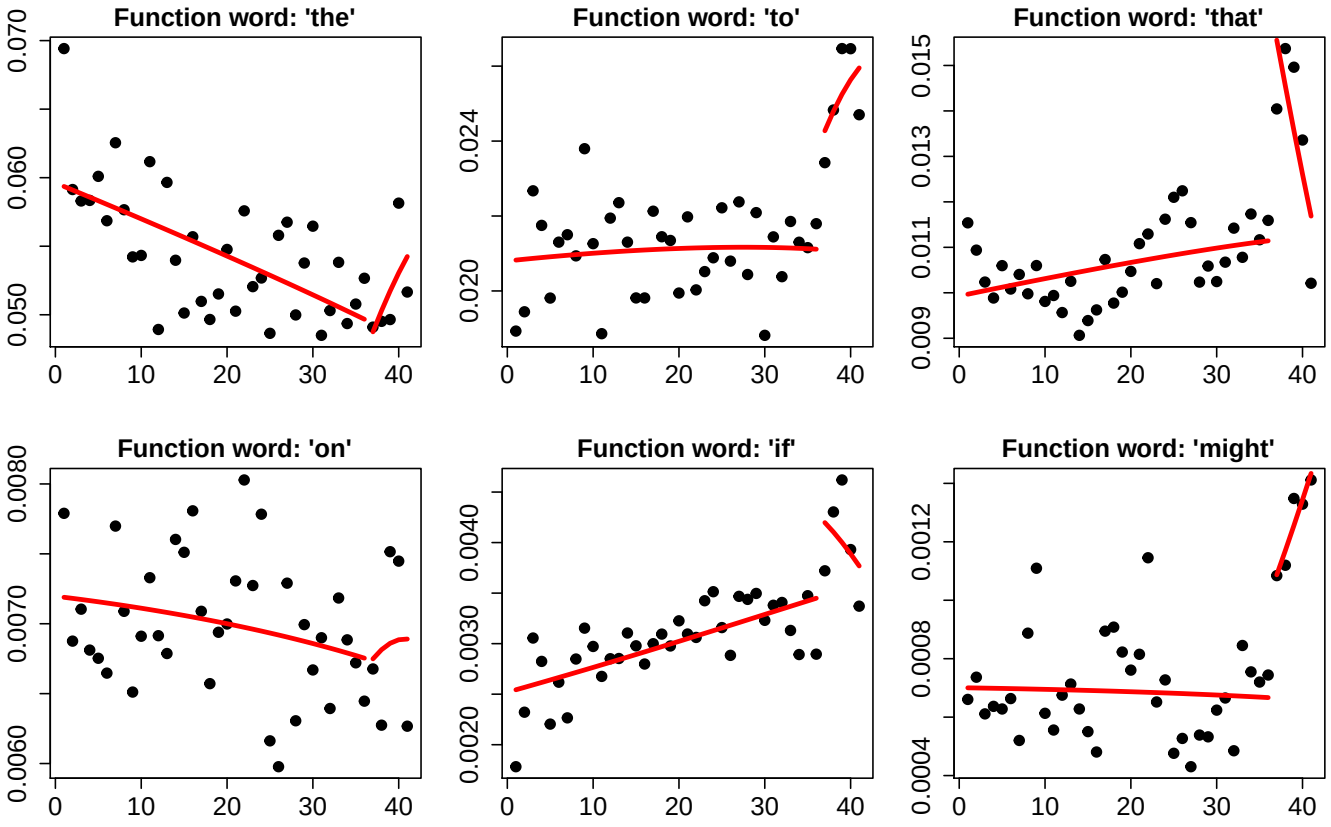

Fig. 6. Evolution of some selected function word proportions over time in the Pratchett corpus, with the best fitting Dirichlet-Multinomial regression lines and change points superimposed.

allows both to occur together. This suggests that the time-evolution of Pratchett's writing style has several aspects, including both slow change as he matured, along with more abrupt changes.

Under the best fitting Dirichlet-Multinomial specification which allows for both gradual drift and abrupt change, there is a single change points detected which occur after the 35th book ('Making Money', written in 2007). To illustrate the fitted model, Figure 6 show plots of a selected number of function word features evolve over time, with the fitted regression lines and change points superimposed. From these, it can be seen that the linear regression change point specification looks reasonable, and that Pratchett's style does indeed seem to undergo both gradual and abrupt change.

From this analysis, we see that the final three non-young adult books which Pratchett wrote after the 2007 change point have a substantially different style to his previous work, even taking into account the gradual drift in style that is occurring over the whole corpus. The first of these three novels is "Unseen Academicals", which was published in 2009. Pratchett's Alzheimer's diagnosis was first made public in December 2007 (BBC News, 2015) suggesting that it would have been present during the writing of this novel. As such, the location of the change point coincides with the Alzheimer's diagnosis, suggesting that this did indeed result in a detectable change in his style.

4.2. Simulation Study

The above analysis shows that the model with both gradual drift and a single change point is favoured for the Pratchett corpus. However there are potential concerns about whether this finding might be a false positive – perhaps the corpus actually does not contain a change point, and the wrong model has been chosen. To address these concerns, we would like to know the probability of incorrectly detecting the existence of drift or abrupt change if the writing style is actually constant over time.

For this purpose, we can use a variant of the permutation test (Higgins, 2004), which is a commonly used resampling technique for hypothesis testing. The Pratchett corpus contains n books $B_1, \ldots, B_n$ ordered in time from the first book to the last. Suppose that we randomly permuted the ordering of these books to create a new corpus $B_{(1)}, \ldots, B_{(n)}$ which consists of the same books but rearranged into a different order. Since this reordering has been done completely at random then there will be no structural drift or change in the new corpus. We can then fit all 8 models to this rearranged corpus and find which one is chosen using the maximum penalised likelihood approach from the previous section. Suppose we then repeat this procedure M times by using M different random orderings, and that in R of these cases a model with drift or abrupt change is selected. Then, R/M is an estimate of the probability that we will incorrectly flag that drift or change has occurred when it truly hasn't. Note that this procedure is essentially a permutation test where the null hypothesis is that no change occurs in writing style over time, with the alternative hypothesis being the presence of drift and/or abrupt change.

To implement this procedure, we chose $M = 10000$ and simulated this many reorderings of the corpus. In only 127 of these cases was a model with drift or abrupt change selected, and the Dirichlet-Multinomial model with no drift or change was selected in the other 9873 cases. As such, the probability of incorrectly rejecting the null hypothesis of no change/drift is approximately 0.012 which suggests that the discovered change point in the previous section is likely to be genuine.

4.3. Agatha Christie

We have also used the methodology above to study the work of another well-known author. Agatha Christie is a prominent English crime fiction writer who published 62 major novels between the years 1920 and 1976. We chose to study her since there is speculation that she suffered from undiagnosed Alzheimer's disease during the last four years of her life, which would have affected the writing of her final three novels. This was previously investigated using quantitative methods by Hirst and Feng (2012) and Le et al. (2011), who found evidence that the writing style of her final books differed from the style of her previous ones and suggested that this may be due to Alzheimer's. However as we have found from the analysis of the Pratchett corpus, it is potentially misleading to directly compare the early and late works of a writer, since any discovered stylistic change might simply be due to the cumulative effects of gradual drift, unconnected to any major life events. Table 10 the Appendix lists all the novels in the Christie corpuses, and each novel was again converted into 201-dimensional feature vector summarising the frequency of function words.

As a preliminary analysis, Figure 7 shows a Multidimensional Scaling plot of the Christie corpus. Recall from Section 2 that MDS projects each book into a two-dimensional feature space, with the distances between each numbered book on the plot corresponding to the distances between their associated feature vectors in the original space. From this plot, we can see evidence of gradual stylistic change over time, similar to that previously observed for the Pratchett corpus in Figure 2. As in the earlier corpus, Christie's successive books are

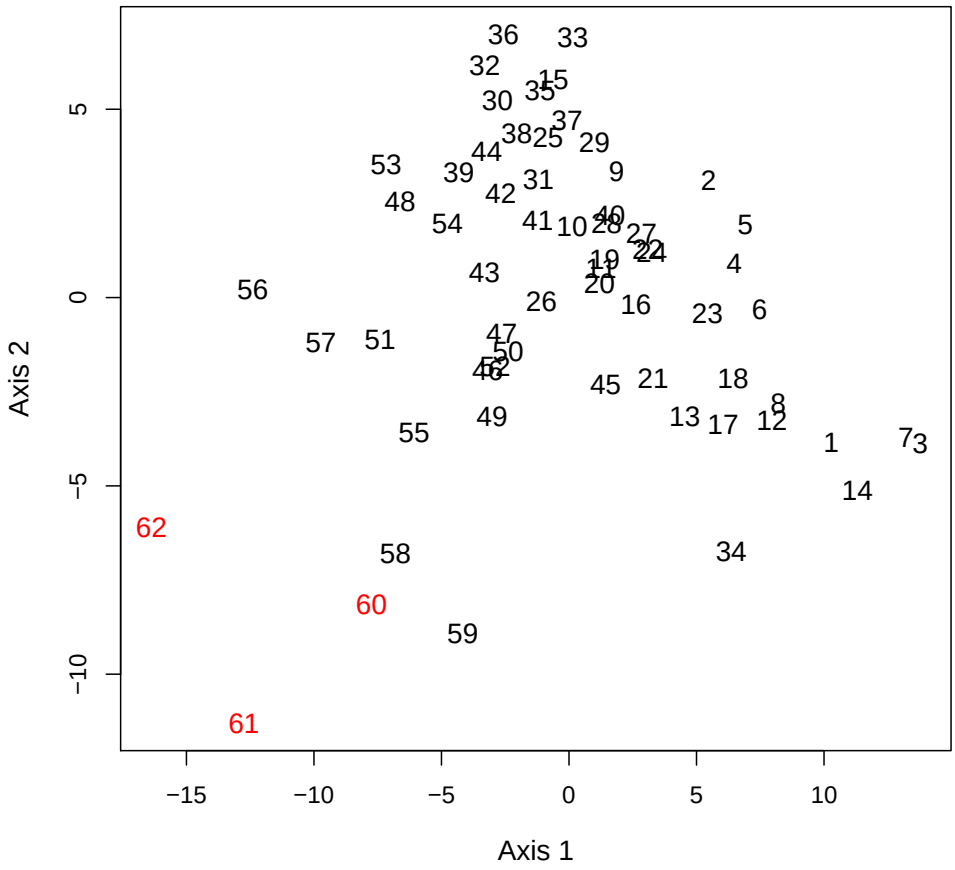

Fig. 7. Multidimensional Scaling plot of the Christie corpus. The novels are numbered in publication order, with the first novel given the number '1', the second given the number '2', and so on. The three novels which were written after Christie's suspected onset of Alzheimer's are highlighted in red.

Fig. 8. Maximum likelihood of model fits on the Christie corpus, penalised by BIC penalty. Larger (less negative) values indicate a more preferable model. The 'Gradual Drift' and 'Drift + Change' entries are indentical because the latter never selected a model containing change points.

Model	Multinomial	Dirichlet-Multinomial
Constant	-115106	-61575
Gradual Drift	-102091	-60890
Abrupt Changes	-88440	-61127
Drift + Changes	-77093	-60894

substantially more similar than books which were written many years apart. The final three novels (60-62) written by Christie are highlighted in red on the left hand plot. Although these do seem to be quite different to her early work (e.g. novels 1-20), they do not seem too dissimilar to the previous 10 novels (50-59) that she wrote prior to her suspected Alzheimer's. This casts preliminary doubt on the claim that the observed difference between her early and late period works is primarily due to Alzheimer's rather than gradual changes in her style over the intervening time period.

To investigate further, we fitted the 8 models for stylistic time-evolution to this corpus, and the resulting BICs are shown in Table 8. As with the Pratchett corpus, there is too much within-corpus variation for the Multinomial distribution to give an adequate fit, suggesting that the compound Dirichlet-Multinomial distribution is necessary. However unlike before, the best fitting model is the one which only has gradual drift rather than any change points. As such, neither the formal statistical analysis nor the visual representations seem to provide evidence that the final novels written by Christie have a substantially different style from the preceding ones, which constitutes some evidence against the claim that Alzheimer's affected her writing. However we cannot rule out the possibility that Alzheimer's might have impacted some stylistic features other than the function words which we have considered

5. Conclusion

The question of whether authorial style changes over time has been relatively unexplored in the stylometric literature, with the typical assumption being that it does not. In this paper we have developed a framework for testing this assumption, and for modelling any changes that may exist. Through the study of two different authors, we have found that writing style does seem to undergo considerable change over time, both gradually and abruptly., even when only considering grammatical features such as the use of function words, This casts doubt on recent claims that authors tend to have a singular 'styleme' that can be identified in all of their writings (van Halteren et al., 2005). This has obvious implications for questions relating to authorship attribution.

A related question concerns the nature of the abrupt changes that can occur over the course of an author's lifetime. We have explored the hypothesis that severe events in an author's life such as the onset of Alzheimer's disease can have an immediate and identifiable impact on writing style. Strong evidence of this was found in the case of Terry Pratchett., with an identified change point occurring around the time when he contracted the disease. However contrary to several other published studies, we did not find a similar change in the work of Agatha Christie. We believe the reason for this is that previous studies have typically studied the change in her style by simply comparing the books she wrote after the suspected Alzheimer's to her early writing. But as we have seen, the presence of gradual drift in authorial style means that the early and late style of an author may differ even though no

abrupt change has occurred. This points towards the importance of distinguishing between this sort of natural change in style, and abrupt changes due to life events.

A potentially interesting issue that we did not consider is the variation of authorial style across genres. As discussed in Section 2.2, the 'Young Adult' Discworld novels seem to have a slightly different style to the books, which is why we excluded them from the analysis. We also did not consider any of the non-Discworld books written by Pratchett, in order to keep the corpus as homogenous as possible. A possible extension of our work would involve the use of covariates or a hierarchal structure to model changes in writing style across multiple genres. Finally it should be noted that our analysis is limiting only to testing whether a change in style has occurred, rather than giving a causal explanation of the specific factors which caused the change. For example, it has been reported that Alzheimer's caused Pratchett to change his method of writing from typing his novels personally, to dictating them to an assistant BBC News (2015). This may be one of the factors responsible for the change in style, but confirming this hypothesis is beyond the scope of the current study.

References

Abakuks, A. (2012) The synoptic problem: on Matthew's and Luke's use of Mark. *Journal of the Royal Statistical Society Series A*, **175**, 959–975.

Agresti, A. (2013) *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley.

Argamon, S. and Levitan, S. (2005) Measuring the usefulness of function words for authorship attribution. *Proceedings of the 2005 ACH/ALLC Conference,*.

BBC News (2015) Obituary: Sir Terry Pratchett [Online]. Available from http://www.bbc.co.uk/news/entertainment-arts-25401679.

Borg, I. and Groenen, P. J. F. (2005) *Modern multidimensional scaling: Theory and applications*. Springer.

Can, F. and Patton, J. M. (2004) Change of writing style with time. *Computers and the Humanities*, **38**, 61–82.

Chen, J., Gupta, A. K. and Pan, J. (2006) Information criterion and change point problem for regular models. *Sankhy: The Indian Journal of Statistics (2003-2007)*, **68**, 252–282.

Chen, J. and Li, H. (2013) Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *Annals of Applied Statistics*, **7**, 418–442.

Doyle, G. and Elkan, C. (2009) Accounting for burstiness in topic models. *Proceedings of the 26th International Conference on Machine Learning*.

Efron, B. and Tibshirani, R. J. (1983) *An Introduction to the Bootstrap*. Chapman & Hall.

Fearnhead, P. (2006) Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, **16**, 203–213.

Fryzlewicz, P. (2014) Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, **42**, 2243–2281. URL: `https://doi.org/10.1214/14-AOS1245`.

Gill, P. S. and Swartz, T. B. (2011) Stylometric analyses using Dirichlet process mixture models. *Journl of Statistical Planning and Inference*, **141**, 3665–3674.

Giron, J., Ginebra, J. and Riba, A. (2005) Bayesian analysis of a multinomial sequence and homogeneity of literary style. *The American Statistician*, **59**, 19–30.

Green, P. (1995) Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **84**, 711–732.

van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M. and Neijt, A. (2005) New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, **12**, 65–77.

Hannart, A. and Naveau, P. (2012) An improved bayesian information criterion for multiple change-point models. *Technometrics*, **54**, 256–268. URL: `http://www.jstor.org/stable/41714894`.

Higgins, J. (2004) *An introduction to modern nonparametric statistics*. Thomson Brooks/Cole.

Hirst, G. and Feng, V. W. (2012) Changes in style in authors with Alzheimer's disease. *English Studies*, **93**, 357–370.

Holmes, D. I. (1985) The analysis of literary style - a review. *Journal of the Royal Statistical Society Series A*, **148**, 328–341.

Juola, P. (2006) Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**, 233–334.

Killick, R., Fearnhead, P. and Eckley, I. (2013) Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107**, 1590–1598.

Koppel, M., Schler, J. and Argamon, S. (2009) Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**, 9–26.

Koppel, M., Schler, J. and Bonchek-Dokow, E. (2007) Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, **8**, 1261–1276.

Le, X., Lancashire, I., Hirst, G. and Jokel, R. (2011) Longitdinal detection of dementia through lexical and syntactic changes in writing: A case studattributionBritish novelists. *Literary and Linguistic Computing*, **26**, 435–461.

Lukashenko, R., Graudina, V. and Grundspenkis, J. (2007) Computer-based plagiarism detection methods and tools: an overview. *Proceedings of the 2007 International Conference on Computer Systems and Technologies*.

Madigan, D., Genkin, A., Lewis, D. D. and Fradkin, D. (2005) Bayesian multinomial logistic regression for author identification. *25th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, **803**, 509–516.

Madsen, R. E., Kauchak, D. and Elkan, C. (2005) Modeling word burstiness using the Dirichlet distribution. *Proceedings of the 2nd International Conference on Machine Learning*.

Menard, S. (2002) *Applied Logistic Regression Analysis*. SAGE Publications, Inc.

Minka, T. (2000) Estimating a Dirichlet distribution. *Tech. rep.*, MIT.

Mosteller, F. and Wallace, D. L. (1963) Inference in an authorship problem. *Journal of the American Statistical Association*, **58**, 275–309.

Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R. and Songn, D. (2012) On the feasibility of internet-scale author identification. *IEEE Symposium on Security and Privacy*.

Pan, J. and Chen, J. (2006) Application of modified information criterion to multiple change point problems. *Journal of Multivariate Analysis*, **97**, 2221–2241.

Pearl, L. and Steyvers, M. (2012) Detecting authorship deception: a supervised machine learning approach using author writeprints. *Literary and Linguistic Computing*, **27**, 183–196.

Peng, R. D. and Hengartner, N. W. (2002) Quantitative analysis of literary styles. *The American Statistician*, **56**, 175–185.

Pierce, D. A. and Schafer, D. W. (1986) Residuals in generalized linear models. *Journal of the American Statistical Association*, **81**, 977–986.

Riba, A. and Ginebra, J. (2006) Change-point estimation in a multinomial sequence and homogeneity of literary style. *Journal of Applied Statistics*, **32**, 61–74.

Ross, G. J., Tasoulis, D. K. and Adams, N. M. (2011) Nonparametric monitoring of data streams for changes in location and scale. *Technometrics*, **53**, 379–389.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Smith, M. W. A. (1983) Recent experience and new developments of methods for the determination of authorship. *Association for Literary and Linguistic Computing Bulletin*, **11**, 73–82.

Thisted, R. and Efron, B. (1987) Did Shakespeare write a newly discovered poem? *Biometrika*, **74**, 445–455.

Uzuner, O., Katz, B. and Nahnsen, T. (2005) Using syntactic information to identify plagiarism. *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, 37–44.

Wallace, H. M. (2006) Topic modelling: beyond bag-of-words. *Proceedings of the 23rd International Conference on Machine Learning*.

Zhang, N. and Siegmund, D. (2007) A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63**, 22–32.

Zhao, Y. and Zobel, J. (2005) Effective and scalable authorship attribution using function words. *Lecture Notes in Computer Science, Berlin, Springer*.

6. Appendix: Description of Corpuses

Fig. 9. List of Discworld novels in Pratchett corpus, along with publication date and Young Adult status

Number	Title	Year	YA	Number	Title	Year	YA
1	The Colour of Magic	1983	No	22	The Last Continent	1998	No
2	The Light Fantastic	1986	No	23	Carpe Jugulum	1998	No
3	Equal Rites	1987	No	24	The Fifth Elephant	1999	No
4	Mort	1987	No	25	The Truth	2000	No
5	Sourcery	1988	No	26	Thief of Time	2001	No
6	Wyrd Sisters	1988	No	27	The Last Hero	2001	No
7	Pyramids	1989	No	28	The Amazing Maurice	2001	Yes
8	Guards! Guards!	1989	No	29	Night Watch	2002	No
9	Eric	1990	No	30	Wee Free Men	2003	Yes
10	Moving Pictures	1990	No	31	Monstrous Regiment	2003	No
11	Reaper Man	1991	No	32	A Hat Full of Sky	2004	Yes
12	Witches Abroad	1991	No	33	Going Postal	2004	No
13	Small Gods	1992	No	34	Thud!	2005	No
14	Lords and Ladies	1992	No	35	Wintersmith	2006	Yes
15	Men at Arms	1993	No	36	Making Money	2007	No
16	Soul Music	1994	No	37	Unseen Academicals	2009	No
17	Interesting Times	1994	No	38	I Shall Wear Midnight	2010	Yes
18	Maskerade	1995	No	39	Snuff	2011	No
19	Feet of Clay	1996	No	40	Raising Steam	2013	No
20	Hogfather	1996	No	41	The Shepherd's Crown	2015	Yes
21	Jingo	1997	No				

Fig. 10. List of novels in the Agatha Christie corpus, along with publication date. The six novels written under the pen name 'Mary Westmacott' are omitted, as are 'Curtain' and 'Sleeping Murder' due to uncertain publication date.

Number	Title	Year	Number	Title	Year
1	Mysterious Affair	1920	32	The Moving Finger	1942
2	The Secret Adversary	1922	33	Towards Zero	1944
3	The Murder on the Links	1923	34	Death Comes as the End	1945
4	The Man in the Brown Suit	1924	35	Sparkling Cyanide	1945
5	The Secret of Chimneys	1925	36	The Hollow	1946
6	Murder of Roger Ackroyd	1926	37	Taken at the Flood	1948
7	Big Four	1927	38	Crooked House	1949
8	Mystery of the Blue Train	1928	39	A Murder is Announced	1950
9	The Seven Dials Mystery	1929	40	They Came to Baghdad	1951
10	The Murder at the Vicarage	1930	41	Mrs McGinty's Dead	1952
11	The Sittaford Mystery	1931	42	They Do It With Mirrors	1952
12	Peril and End House	1932	43	After the Funeral	1953
13	Lord Edgware Dies	1933	44	A Pocket Full of Rye	1953
14	Murder on the Orient Express	1934	45	Destination Unknown	1954
15	Why Didn't They Ask Evans	1934	46	Hickory Dickory Dock	1955
16	Three Act Tragedy	1935	47	4:50 From Paddington	1957
17	Death in the Clouds	1935	48	Ordeal By Innocence	1958
18	The ABC Murders	1936	49	Cat Among the Pigeons	1959
19	Murder in Mesopotamia	1936	50	The Pale Horse	1961
20	Cards on the Table	1936	51	The Mirror Crack'd from Side to Side	1962
21	Dumb Witness	1937	52	The Clocks	1963
22	Death on the Nile	1937	53	A Caribbean Mystery	1964
23	Appointment With Death	1938	54	At Betrram's Hotel	1965
24	Hercule Poirot's Christmas	1938	55	Third Girl	1966
25	Murder is Easy	1939	56	Endless Night	1967
26	Sad Cypress	1940	57	By the Pricking of my Thumbs	1968
27	One Two Buckle My Shoe	1940	58	Halloween Party	1969
28	Evil Under The Sun	1941	59	Passenger to Frankfurt	1970
29	N or M	1941	60	Nemesis	1971
30	The Body in the Library	1942	61	Elephants Can Remember	1972
31	Murder in Retrospect	1942	62	Postern of Fate	1973